INL/CON-10-18552
PREPRINT

Lessons Learned on Benchmarking from the International Human Reliability Analysis Empirical Study

PSAM 10

Ronald L. Boring
John A. Forester
Andreas Bye
Vinh N. Dang
Erasmia Lois

June 2010

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

Lessons Learned on Benchmarking from the International Human Reliability Analysis Empirical Study

Ronald L. Boring,[a*] John A. Forester,[b] Andreas Bye,[c] Vinh N. Dang,[d] Erasmia Lois[e]
[a]Idaho National Laboratory, Idaho Falls, Idaho, USA
[b]Sandia National Laboratories, Albuquerque, New Mexico, USA
[c]OECD Halden Reactor Project, Halden, Norway
[d]Paul Scherrer Institute, Villigen PSI, Switzerland
[e]U.S. Nuclear Regulatory Commission, Washington, DC, USA

Abstract: The International Human Reliability Analysis (HRA) Empirical Study is a comparative benchmark of the prediction of HRA methods to the performance of nuclear power plant crews in a control room simulator. There are a number of unique aspects to the present study that distinguish it from previous HRA benchmarks, most notably the emphasis on a method-to-data comparison instead of a method-to-method comparison. This paper reviews seven lessons learned about HRA benchmarking from conducting the study: (1) the dual purposes of the study afforded by joining another HRA study; (2) the importance of comparing not only quantitative but also qualitative aspects of HRA; (3) consideration of both negative and positive drivers on crew performance; (4) a relatively large sample size of crews; (5) the use of multiple methods and scenarios to provide a well-rounded view of HRA performance; (6) the importance of clearly defined human failure events; and (7) the use of a common comparison language to "translate" the results of different HRA methods. These seven lessons learned highlight how the present study can serve as a useful template for future benchmarking studies.

Keywords: HRA, benchmark, lessons learned

1. INTRODUCTION

1.1. Background

Since the advent of a Technique for Human Error Prediction (THERP) for nuclear power applications in 1983 [1], there has been continuous development and refinement of methods in human reliability analysis (HRA). A recent survey [2] suggests there may now be as many as 72 HRA methods in various guises, although the number of fully implemented and used methods is estimated to be around 35. Of these, the number frequently used, especially in the nuclear industry, is considerably smaller. HRA methods differ along a number of dimensions, including their scope, underlying model, underlying data, and approach to quantification [3]. HRA methods have been further classified according to first and second generation [4-7]; task-, time-, or context-related [8]; or atomistic and holistic [9].

Because of the wide variety of HRA methods, it is important to compare them to determine their overlap and differences, especially in terms of their analysis outcomes. A number of expert comparisons have been conducted that evaluate HRA methods according to subjective criteria [2-3, 10-12], but there have been very few actual HRA benchmark activities that have compared the outputs of various HRA methods. A *benchmark* in conventional language use requires a reference or standard to which something can be compared. Benchmarking in the present context refers both to comparing HRA methods to each other and to an objective empirical reference such as operating crew performance in a nuclear power plant simulator. Thus, HRA benchmarking can involve both the validation of method predictions against standard performance (i.e., do the methods make accurate

[*]Address correspondence to Ronald Laurids Boring, PhD, Human Factors, Controls, and Statistics Department, Idaho National Laboratory, Idaho Falls, ID 83415, USA. Email: ronald.boring@inl.gov.

predictions?) and against each other in terms of their ability to accurately predict and explain the basis for their predictions (i.e., how do the methods compare to one another?).

A recent literature review [13] discusses the HRA benchmarks that have been conducted to date, including:

- The large-scale Human Factors Reliability Benchmark Exercise (HF-REB) [14] conducted at the Joint Research Center of the European Commission in Ispra, Italy, in the late 1980s, focusing on three nuclear power plant scenarios.
- Zimolong's [15] study to validate three HRA method predictions against actual human performance on a simulated batch manufacturing scenario.
- Kirwan's [16-18] benchmark of 30 analysts applying three HRA methods widely used in the British nuclear industry.
- Maguire's [19] extensive validation of predictions by one HRA method to operational data from aviation.

The review of previous HRA benchmarks [13] highlights a number of lessons learned. In the case of the validation benchmarks, the emphasis was on comparing the HRA method predictions to empirical data. In other cases, the emphasis was on a method-to-method benchmark, in which method results were compared to each other but not to an external standard. The previous benchmark studies showed that there was considerable variability in HRA method predictions. This variability can be attributed to variability in the application of individual methods or to the ability of the methods to handle a wide variety of scenarios. Variability in the application of individual methods can be attributed to a lack of clear definition or understanding of the scenario being analyzed—a byproduct of the benchmark study design. Additionally, previous benchmarks focused heavily on one aspect of the analysis—the quantitative output produced by the application of the method. Additional insights into the process behind the analysis—including key assumptions made for the qualitative side of HRA—would have been useful for explaining possible sources of variability.

Another important lesson learned from earlier studies concerns the infrequent nature of error in skilled performance. For those studies involving empirical operator data, it becomes important to acknowledge this low frequency and compensate either with a large number of participants for between-subject study designs or a large number of test runs for within-subject study designs. If such manipulations are impractical, it is possible in many cases to seed error-prone behavior, e.g., by increasing the workload or task complexity. Such error seeding must be carefully controlled so as to avoid confounds or artificiality, but challenging situations can and do occur in most human operated systems.

1.2. The International HRA Empirical Study

Since the last large-scale HRA benchmark at Ispa [14], a considerable number of new HRA methods has been developed, and HRA has grown in its application both within and outside the nuclear industry. Moreover, the Ispra HF-RBE consisted of a method-to-method comparison, while empirical validation benchmark studies [15, 19] have generated an interest to translate the method-to-data approach into a larger scale study. Consequently, an international group of HRA researchers, led by the U.S. Nuclear Regulatory Commission and the Organization for Economic Co-Operation and Development (OECD) Halden Reactor Project, formulated a new HRA benchmark called the "International HRA Empirical Study" [20-22]. This study was designed to build upon earlier HRA benchmark efforts, specifically by looking at both qualitative and quantitative analyses, holding constant the information provided to different analysis teams, providing information about (but not direct access to) the crews in the study, using performance shaping factors to allow comparison of degraded but not outright failed operator performance, and providing a common template to allow ready comparison of predictions across methods and from methods to empirical data [23].

Figure 1 depicts the design of the International HRA Empirical Study. Fourteen nuclear power plant (NPP) crews participated in the study at the Halden huMan-Machine Laboratory (HAMMLAB), a full-scope NPP control room simulator. These crews consisted of a licensed reactor operator, an assistant reactor/turbine operator, and a shift supervisor. Each crew participated in two simulated scenarios—a steam generator tube rupture (SGTR) and a loss of feedwater (LOFW) incident. Both scenarios had a simple (or base) case that closely approximated how such an incident might be trained for during routine simulator training. In addition, both scenarios had a complex case, in which the familiar SGTR or LOFW scenario was complicated by secondary malfunctions. Crew performance was observed and documented during these scenario runs in the simulator. The crew performance documentation consisted primarily of a standardized set of drivers based on the performance shaping factors in NUREG-1792 [24], short operational summaries of the crew actions, and the success or failure of the crews to complete specific actions, sometimes within a predefined time window. The specific actions corresponded to human failure events (HFEs) that are defined in probabilistic risk assessment (PRA) models.[1]

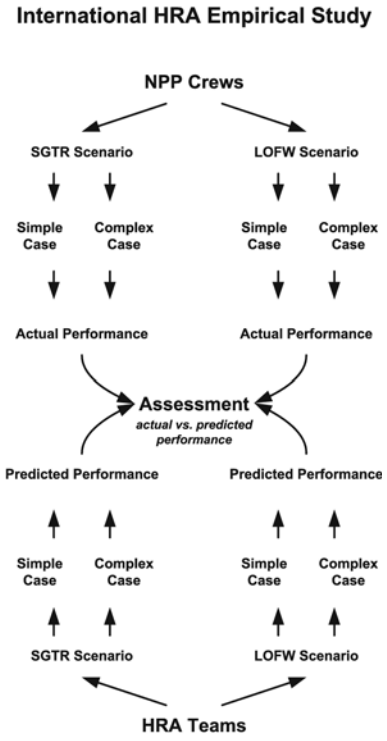

Figure 1: The Design of the International HRA Empirical Study

Fourteen HRAs applied with 13 different HRA methods were completed (one method was used by two analysis teams), each generating unique drivers, operational stories, and human error probabilities (HEPs). The outputs produced by the teams were comparable to the performance documented for the HAMMLAB crews, thus allowing a direct comparison between the empirical data and the predicted

[1] Note that the HFEs were defined in some cases according to whether or not crews were able to complete the tasks within a specified time as expected by their training. This performance criterion alone is not standard in PRAs, which define an HFE as a failure of a function, system, or component that is the result of human actions or inactions [25]. The additional time criterion was included based on trainer expectations of good performance in order to define HFEs that might challenge crews and increase the probability of failure.

HRA data by the assessment team. Note that because the crews had highly reliable performance in most cases [23], it was not always possible to generate failure rates directly from the crew performance data. As such, the benchmark focused more on comparing qualitative data (i.e., the drivers and the operational stories) than quantitative data (i.e., the actual crew failure rates vs. the methods' predicted HEPs).

2. Lessons Learned

The International HRA Empirical Study featured a number of important lessons learned about conducting a benchmarking activity. The careful design of the study as a benchmark was key to ensuring HRA methods were effectively evaluated against actual performance and compared on meaningful criteria. The following sections highlight seven of the most significant lessons learned about HRA method benchmarking from the study. Additional lessons learned about the HRA methods themselves can be found in the companion piece, [26].

2.1. Study Design Issues

The International HRA Empirical Study was merged with another HRA study at the HAMMLAB [27]. The objective of that study was to manipulate the complexity of the tasks performed by control room crews, thus allowing a direct comparison of crew performance on simple vs. complex tasks. Early in the task complexity study design process, many persons who later joined the benchmark assessment group met with the HAMMLAB group in order to discuss scenarios. The purpose was to ensure that the HAMMLAB study used PRA relevant scenarios and in general was fit for HRA purposes. Based on these considerations, we knew that these scenarios were also a good fit for HRA benchmarking purposes. Because the overall study was a merger of task complexity and benchmarking objectives, it can be seen as a successful dual-purpose study.

An advantage of the dual-purpose study is that it eliminated many of the barriers to running a standalone benchmarking study. Foremost among these barriers are the resources required to support dedicated simulator runs in addition to organizing analysis teams to perform the HRAs. By merging the benchmark with the task complexity study, resources were optimized.

In the benchmark study, separate groups were organized to perform respective tasks:

- The HAMMLAB group was responsible for designing and conducting the empirical data part of the study, including observing and summarizing the crew performance.
- Independent analysis teams performed HRAs on the scenarios to predict crew performance.
- The HRA assessment team was responsible for comparing the actual crew performance findings to the HRA predictions.

The activities were separated, facilitating a true "blind" comparison of the findings with the predictions. HRA predictions were summarized and put into a common template to allow easy comparison (see Section 2.7). This summary was done without the assessment team being aware of the actual findings from the crews. After the summaries were completed, the assessment team learned of the actual crew performance, and comparisons were made between the findings and predictions. This division of labor helped the assessment team to perform the comparisons objectively. In addition, the data analysis team in Halden did not see any HRA analyses from the HRA teams before they had finished the analysis of the empirical data. These information firewalls helped avoid any impact from one part of the study to another.

A minor artifact of the dual nature of the study concerned the timing of the study. Because the analysis teams had not yet been finalized at the time the study was run, it was not possible to have the teams observe real crews in the simulator, nor interview crews to answer questions related to factors like crew familiarity with the scenario or key aspects of crew operational culture. It is a standard part of HRA to allow the analysis teams to watch simulator runs and ask questions of representative crews

directly [24, 28-29]. In order to compensate for the HRA teams' lack of such hands-on information, the Halden group and the benchmark assessment group developed an extensive information package in the study. This package comprised the emergency operating procedures used in HAMMLAB; a detailed description of the simulator environment, including a video showing Halden staff acting as a crew in a reference simulation scenario; and detailed descriptions of the scenarios, including alarm lists and screen-shots from the simulator screens showing key indications at different times of the events. Additionally, there was information on the crews and their typical characteristics. The information package was followed up by questions and responses from individual HRA teams about certain aspects of the scenarios or simulator environment. All questions and responses were circulated to all teams. The advantage of this information package and the question-and-answer round was that all HRA teams received exactly the same information. This facilitated control of the analysts' crew and plant knowledge, albeit it might be argued that some HRA methods include mechanisms for data collection that others don't, and some methods need more detailed information than do other methods.

2.2. Benchmark to Empirical Qualitative and Quantitative Data

As noted, this study consists mainly of a method-to-data benchmark comparison, with the main emphasis on how the individual methods matched the empirical data, rather than on how methods compared to other methods. This benchmarking approach deemphasizes trying to determine whether one method is comparatively better or worse than another. The premise of the current study is that each HRA method was designed predicated on different assumptions and is optimized to different applications. Thus, a single benchmark is inadequate to rank the relative merits of HRA methods for all conditions/applications. Instead, the benchmark serves as a way to evaluate the HRA methods individually in their predictive efficacy for a specific application, but does allow some insights regarding strengths and weaknesses of the method.

HRA consists of qualitative and quantitative phases. Once HFEs have been identified, they are first analyzed in terms of those factors that might contribute to the performance failure likelihood. In most HRA methods, this qualitative phase consists of evaluating the context and drivers on performance—typically in the form of performance shaping factors (PSFs). After the context and drivers are determined, this information is used by the analysts to quantify the HEP. HRA methods have distinct approaches to completing the qualitative and quantitative portions of the analyses. Some methods, such as those associated with root cause analysis, are primarily qualitative. Other methods, such as many of the simplified HRA approaches, do not provide a formal qualitative approach and instead only provide a means to quantify the HFE. For the simplified approaches, it is nonetheless assumed that a qualitative analysis will be performed by the analysts prior to quantification. In practice, the use of different qualitative approaches may lead to considerable variability in those HRA methods that are primarily quantitative, because two analyses using the same quantification approach may make use of different qualitative approaches.

An important lesson learned from the International HRA Empirical Study is the value of considering not just quantitative findings in comparing methods to crew performance. Rather, the study sought to capture qualitative factors—both context and drivers on performance—from the crews and the methods. Performing only a quantitative comparison is analogous to doing a math problem without showing the work behind the answer. If the quantitative result doesn't match the actual data, does this mean the HRA method represents a poor analysis tool? The study revealed cases where HRA methods performed a thorough qualitative analysis but did not match the empirical quantitative findings, suggesting a potential poor mapping between the methods' qualitative and quantitative phases. In other cases, methods identified incorrect drivers in their qualitative analysis but had good matches between their HEPs and the error rates observed in the operating crews. In such cases, it might be argued that the HRA method arrived at the right quantitative result but for the wrong reasons. Whatever the case, the inclusion of both qualitative and quantitative elements in the comparison allowed a more complete understanding of the HRA methods' strengths and weaknesses in predicting crew performance.

2.3. Consideration of Negative and Positive Drivers

The previous section noted the importance of performing a qualitative comparison in the International HRA Empirical Study. A closely related topic is the consideration of both negative and positive drivers on human performance. Early HRA methods tended to weight only the deleterious effect of PSFs on performance [30]. Such methods might be seen as subtractive—quantification starts with the assumption of nominal performance, whereupon factors that might negatively impact performance would be considered and subtracted from the nominal level of human reliability. Of course, since HRA primarily operates in failure space, the subtractive treatment has the effect of increasing the HEP—as reliability decreases, the HEP increases. Many contemporary HRA methods also consider the positive effects of PSFs on performance by crediting those factors that might increase the likelihood of success. Such methods might be seen as additive—success-enhancing factors are added to the nominal performance, resulting in increased reliability and a lower HEP.

The practical implication of negative and positive PSFs is seen most directly in HRA quantification. However, the International HRA Empirical Study revealed the value of such categorization also in terms of qualitative insights and comparisons. While not all HRA methods incorporate both negative and positive PSFs, the number was significant enough to warrant the inclusion of both poles for the drivers used in documenting crew performance. In fact, on average across all SGTR scenarios and HFEs, a slight positive effect was credited for these drivers: Execution Complexity (meaning the difficulty to carry out the task, not diagnose the right course of action), Procedural Guidance, the Human-Machine Interface, Work Processes, and Communication [22]. This means, on average, evidence suggested that these factors contributed to crew success across the simulator runs. Individual HRA methods may or may not have been sensitive to the positive effects of such PSFs, but there was value in comparing all methods according to these drivers. For example, if a method predicted a negative effect for a PSF related to Execution Complexity, the disparity between the method predictions and actual performance warranted closer study for the reason behind the disagreement. It must be noted that some negative or positive effects may be difficult to predict for the analysis teams. For example, a driver such as Communication, which was observed to be mostly positive across the SGTR simulator runs, was noted by analysis teams to be difficult to anticipate given the lack of experience the analysis teams had interacting with the specific crews used in the HAMMLAB studies. Despite such limitations, this approach of considering both negative and positive PSFs allowed a more complete understanding of crew successes and shortcomings than was present in earlier benchmark activities, which had focused exclusively on negative drivers.

2.4. Sample Size

The International HRA Empirical Study featured a relatively large number of crews (14 total for each condition), making it possible to study crew-to-crew variability and, to a limited extent, to estimate the human error rate for HFEs within the scenarios. This is especially noteworthy in that the crews represented highly skilled control room operators, each comprising a team of the assistant reactor operator, reactor operator, and shift supervisor. A study of this scale for HRA had not previously been attempted and represents a positive byproduct of being an adjunct to an existing HRA study.

Having multiple crews allows the observation of possible differences in response paths across crews. In fact, the crew operational performance stories provided in [22] highlight clusters of behaviors by crews in response to transients. While the crews stayed within procedural guidance, there was some variability in their responses depending on how the scenario evolved and how they interpreted the situation at hand. Previous research in usability and perception [31-32] suggests that five participants (or participant groups) marks the point at which a study can expect to find approximately 85% of errors that might be encountered for a particular interface or situation. As the number of participants increases, the percentage of errors detected asymptotes to a point of diminishing returns for each added participant. Nonetheless, further research [33] suggests that where individual differences exist, there may be advantages in increasing the number of participants. The International HRA Empirical Study demonstrated the value of having a large sample size, in that crew-to-crew variability was

identified. This crew-to-crew variability was not unbounded. In all cases, clusters of crew responses emerged, suggesting consistency within crews depending on their diagnosis and subsequent treatment of the situation.

The relatively large sample size serves primarily as an advantage to gaining qualitative insights. A representative HRA method [34] suggests that one would expect experienced crews to commit an error 1 in 100 times while performing a diagnosis task and 1 in 1,000 times while performing an action task. These nominal error rates are in line with most HRA methods. The problem with running an HRA study designed primarily for quantitative insights is that it would be necessary to run a large sample of participants before one would reasonably expect to see the occurrence of an error. In the present study, task complexity was intentionally manipulated in order to see the effects of increased complexity on crew performance. As such, the complex cases of the study had predicted error rates in excess of the nominal error rates. In fact, when using the timing criteria (see Footnote 1) to determine success or failure on particular HFEs, some crews did fail at some HFEs in the complex cases. Timing was not the only contributor to failure. Some failure is not unexpected given the demands of the situation and the tight timing criterion. For these types of HFEs and with 14 crews, such observed performance data may be used to arrive at an objective error rate for the crews. The Bayesian analysis performed as part of the analysis shows that the uncertainty bounds are rather narrow when a number of crews fail an HFE [35]. It might be argued that these HFEs are not representative of typical PRAs. The standard PRA HFEs, as in the base case in this study, feature infrequent failure and large uncertainty bounds. In such cases, when the frequency of errors among crews is low, it would take considerable crew runs in the simulator before there would be sufficient statistical power to arrive at a definitive error rate.

2.5. Multiple Methods and Multiple Scenarios

While subjective benchmarks of HRA have compared a large number of methods [2-3, 10-12], the empirical benchmarks discussed in [13] featured a much smaller collection of methods. It was desirable to include a wider number of HRA methods than had been found in previous HRA empirical benchmarks. As such, the International HRA Empirical Study featured 13 different HRA methods performed by 14 analysis teams (one method had two teams). This specific number of methods was not directly by design but rather the happy byproduct of the willingness of many analysis teams to participate in the study. The study design was flexible with respect to how many methods could or should participate. This flexibility resulted in an empirical comparison that featured as many methods as several of the previous subjective benchmarks. This wide-net approach provided a solid cross-section of methods currently used internationally in HRAs at nuclear power plants.

One possible limitation of the current design is that all HRA methods (with one exception) in the study featured only a single analysis team. As such, it is difficult to generalize the comparison findings to the method in general. In the worst case, the comparison findings may reflect the peculiarities of one analysis team and not prove representative of other applications of the same method. Since there is no way to gauge inter-analyst intra-method variability given the makeup of the analysis teams in the present study, future HRA benchmarking efforts should attempt to provide more than one analysis team per method. It is not, however, felt that this limitation hindered successful insights into the methods in this study. There were considerable lessons learned about the methods [26] in terms of their utility for qualitative and quantitative predictions. Additional insights on the process were documented formally and anecdotally by the analysis teams, allowing the study's assessment team to provide informed discussions on strengths and weaknesses of the individual methods with respect to their use in the International HRA Empirical Study. Further generalization was not attempted nor warranted.

Just as one must be cautious to generalize the results of one team's analysis to an entire HRA method, one must take care to consider the specific scenarios that are being analyzed. The present study provided two scenarios—an SGTR and LOFW scenario, along with base and complex case variants—to allow a representative sample of the types of analyses for which the HRA methods might be used.

HRA methods were designed for different purposes, and no single scenario is sufficient to gauge the merits or limitations of a particular method. The two scenarios covered in the present study provided a good starting point for evaluating HRA methods, but these two scenarios are by no means exhaustive. Additional scenarios to span the gamut of HRA activities are a logical extension of the current study.

2.6. Clearly Defined Human Failure Events

The HFE represents the human activities that are included in the PRA model. The HFE is the unit that is quantified as an HEP. While an HFE may be incorporated as a simple node in a fault tree or a branch in an event tree, the documentation supporting the HFE represents the nexus of qualitative insights used during the quantification process. These insights may be simple to detailed, depending on the analysis needs and the level of task decomposition.

HRA methods do not have a consistent level of task decomposition. This lack of consistency can result not only in different qualitative analyses but also different HEPs. Moreover, the level of task decomposition affects the dependency between tasks, which may have a further effect in driving the HEP. The issue is not that different HRA methods necessarily produce different results for the same HFE; rather, different HRA methods may decompose the HFE to different levels. Thus, the quantification of the same HFE may entail different assumptions and, to some extent, different groupings of tasks across HRA methods. In other words, because of a lack of a common task decomposition framework, HRA methods may not be using the same unit of analysis when producing the HEP.

The Ispra HF-RBE [14] demonstrates how central this topic is to HRA. The benchmark featured three phases of analysis to compare HRA methods. Each successive phase served to further bound the HFE. The first phase asked the HRA teams to identify and quantify HFEs. Because different HFEs were identified across methods, it was difficult to compare method results directly. The second phase involved a more explicit definition of the HFEs to ensure the analysis teams quantified the same HFE. Even with a commonly defined HFE, there was considerable variability in how analysis teams modeled the HFE. Differences in task decomposition played a significant role in the differences of the HEPs for the HFEs. Some analysis teams decomposed to a finer level, resulting in lower HEPs. However, the dependencies between HFEs were not well accounted for in the analyses with finer grained task decomposition, resulting in unrealistically low HEP values in the original author's opinion. As such, a third phase was conducted, this time with an explicit decomposition of tasks and a common HRA event tree used in quantification.

In the present study, the HFEs were carefully pre-defined, and clear definitions of the HFEs were provided to all HRA teams to ensure consistent analyses. Providing pre-defined HFEs did not preclude the need for a qualitative analysis by each analysis team. In fact, in contrast to the first two phases of the HF-RBE [14], the present study demonstrated consistent analyses across different teams. The only issue related to the definition of the HFEs in some cases concerned the use of somewhat artificial timing windows for success/failure designation. Because no objective, published standard exists for the proper timing of tasks, there was some variability in the amount of time that the analysis teams predicted each task might take. Much of this variability stemmed from a lack of hands-on familiarity of the analysis teams with the crews and the simulator configuration. This issue manifested in the first series of analyses related to the SGTR scenario. By the second phase of the study, involving the LOFW scenario, analysis teams had been briefed on the performance of the crews during the SGTR scenario compared to the method predictions. Analysis teams commented that this debriefing served as a helpful calibration of their analyses to actual performance. Consequently, in the later LOFW analyses, the assessment team did not observe significant misgauging of the time required by the crews.

2.7. Common Comparison Template

In the original pilot phase of the study documented in [20-21], corresponding to a single HFE each

from the base and complex SGTR scenarios, three types of qualitative data were collected for both the simulator runs and the HRAs:

1. The method-specific PSFs,
2. Operational expressions—narratives to provide the context of how the operator actions unfolded, and
3. A comprehensive error taxonomy.

The comprehensive taxonomy [36] was based on the Human Event Repository and Analysis (HERA) database system [37-38]. Modifications included the elimination of factors irrelevant to the control room simulator study like maintenance factors, balance-of-plant operations, and PSFs related to the environment or fitness for duty. The goal of using the HERA-like taxonomy was to allow a very detailed comparison of the factors that influenced the analysis in the application of the HRA methods. The identical taxonomy was completed for the simulator data, thus allowing a precise comparison of simulator data and HRA method predictions. Feedback from the HRA teams on using the HERA-like taxonomy was mixed [36]. While several teams found the taxonomy useful as a tool to express their analysis in a standard format, others found that the taxonomy was not a perfect match to their HRA method and preferred the more open-ended feedback of providing the PSFs and operational expressions. The taxonomy also proved time consuming for the HRA teams to complete. The time required to complete the taxonomy is a direct reflection of the number of items in the taxonomy. It is important to note that the taxonomy was designed for retrospective analyses, in which case the taxonomy serves as a reference checklist of factors that might have been observed or documented in an event report. The process of considering each item in the taxonomy for a prospective analysis like that conducted by the HRA teams adds considerably to the coding requirements and is an application for which the taxonomy is not yet optimized [36].

By the second and subsequent phases of the study [22], the HERA-like taxonomy was eliminated from the comparison. Where the goal had been for the HERA-like taxonomy to provide a common language by which to make direct method-to-data comparisons possible, this goal was actually adequately served by the other qualitative data. In fact, when the method-specific PSFs were standardized through expert review to the same drivers used to account for crew performance, these drivers became the common language that facilitated the qualitative comparison.

The drivers provided, in a simple form, a roadmap of the thinking behind the analyses and of the factors that influenced the crews' performance. By weighting the drivers according to a negative, neutral, or positive influence, it was possible to see where a method might have systematically under- or over-considered the effects of particular drivers compared to actual crew performance. It was also possible to see where a method might not address particular drivers, or where a method considered drivers that were not captured through the observations. The drivers provided a level of detail that was well suited for an overview of the method-to-data comparison. A more nuanced account was subsequently found in the operational expression, although it was not possible to translate this freeform narrative into a common language. Thus, a two-tiered comparison strategy emerged: the high-level overview of similarities and differences between the analyses and crew performance was possible through the common drivers; a more detailed comparison that was sensitive to the unique aspects of the HRA methods and analysis teams was possible through the operational expressions.

3. Conclusions

Since the data analysis is still ongoing, the above lessons learned on HRA benchmarking represent only a snapshot of the International HRA Empirical Study. These lessons learned nonetheless serve as a template for how benchmarking studies can be run effectively. The lessons learned provide guidance pertaining to a benchmark in practice (e.g., information to teams, blind study), effective study design (e.g., multiple methods and scenarios), and effective analysis (e.g., consideration of positive and negative drivers). Several of these lessons were learned during the study design phase, but additional refinements were realized during the course of the study (e.g., the optimization of the

common comparison language). With the inclusion of these lessons learned, the International HRA Empirical Study marked an improvement over previous benchmarking efforts, leading to greater insights into the HRA methods themselves [26]. Further application and refinement of these lessons learned will fortify the ability of HRA benchmarks to draw useful conclusions. Moreover, by simplifying the process of HRA benchmarking, the lessons learned provide the roadmap for an ongoing empirical research program in HRA.

Acknowledgements and Disclaimers

The authors gratefully acknowledge the many participants in this study, including the control room crews, the simulator observation team, the HRA teams, the assessment team, and independent reviewers.

This study is a collaborative effort of the Joint Programme of the OECD Halden Reactor Project and, in particular, Halden's signatory organizations who provided the HRA teams, the United States Nuclear Regulatory Commission (USNRC), the Swiss Federal Nuclear Inspectorate (DIS-Vertrag Nr. 82610), and the U.S. Electric Power Research Institute. In addition, parts of this work were performed at Sandia National Laboratories and Idaho National Laboratory (INL) with funding from the USNRC. Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000. INL is a multi-program laboratory operated by Battelle Energy Alliance LLC, for the United States Department of Energy under Contract DE-AC07-05ID14517.

References

[1] A.D. Swain and H.E. Guttman, *"Handbook of human reliability analysis with emphasis on nuclear power plant applications, Final report, NUREG/CR-1278,"* US Nuclear Regulatory Commission, 1983, Washington, DC.

[2] J. Bell and J. Holyroyd, *"Review of human reliability assessment methods, RR679,"* Health and Safety Executive, 2009, Buxton, UK.

[3] J. Forester, A. Kolaczkowski, E. Lois, and D. Kelly, *"Evaluation of human reliability analysis methods against good practices, Final report, NUREG-1842,"* US Nuclear Regulatory Commission, 2006, Washington, DC.

[4] E.M. Dougherty, *"Human reliability analysis—Where shouldst thou turn?"* Reliability Engineering and System Safety, 29, pp. 283-299, (1990).

[5] E. Hollnagel, *"Cognitive reliability and error analysis method,"* Elsevier, 1998, Oxford.

[6] US Nuclear Regulatory Commission, *"Technical basis and implementation guidelines for a technique for human event analysis (ATHEANA), NUREG-1624,"* US Nuclear Regulatory Commission, 2000, Washington, DC.

[7] R.L. Boring, *"Dynamic human reliability analysis: Benefits and challenges of simulating human performance,"* in T. Aven & J.E. Vinnem (Eds.), Risk, Reliability and Societal Safety, Volume 2: Thematic Topics. Proceedings of the European Safety and Reliability Conference, ESREL 2007, pp. 1043-1049, Taylor & Francis, 2007, London.

[8] A.J. Spurgin, *"Human reliability assessment theory and practice,"* CRC Press, 2009, Boca Raton.

[9] R.L. Boring and D.I. Gertman, *"Atomistic and holistic approaches to human reliability analysis in the US power industry,"* Journal of the Safety and Reliability Society, 25(2), pp. 21-37, (2005).

[10] A. Swain, *"Comparative evaluation of methods for human reliability analysis, GRS-71,"* Gesellschaft für Reaktorsicherheit, 1989, Cologne.

[11] F.T. Chander, Y.H. Chang, A. Mosleh, J.L. Marble, R.L. Boring, and D.I. Gertman, *"Human reliability analysis methods: Selection guidance for NASA,"* NASA Office of Safety and Mission Assurance, 2006, Washington, DC.

[12] S. Adhikari, C. Bayley, T. Bedford, J. Busby, A. Cliffe, G. Devgun, M. Eid, S. French, R. Keshvala, S. Pollard, E. Soane, D. Tracy, and S. Wu, *"Human reliability analysis: A review and critique, Final report of the EPSRC funded project rethinking human reliability analysis methodologies,"* Manchester Business School, 2008, Manchester, UK.

[13] R.L. Boring, S.M.L. Hendrickson, J.A. Forester, T.Q. Tran, and E. Lois, *"Issues in benchmarking human reliability methods: A literature review,"* Reliability Engineering and System Safety, 95, pp. 591-605 (2010).

[14] A. Poucet, *"Human factors reliability benchmark exercise, Final report, EUR 12222,"* CEC-JRC, 1989, Ispra.

[15] B. Zimolong, *"Empirical evaluation of THERP, SLIM, and ranking to estimate HEPs,"* Reliability Engineering and System Safety, 35, pp. 1-11, (1992).

[16] B. Kirwan, *"The validation of three human reliability quantification techniques—THERP, HEART, and JHEDI: Part I—Technique descriptions and validation issues,"* Applied Ergonomics, 27, pp. 359-373, (1996).

[17] B. Kirwan, *"The validation of three human reliability quantification techniques—THERP, HEART, and JHEDI: Part II—Results of validation exercise,"* Applied Ergonomics, 28, pp. 17-25, (1997).

[18] B. Kirwan, *"The validation of three human reliability quantification techniques—THERP, HEART, and JHEDI: Part III—Practical aspects of the usage of the techniques,"* Applied Ergonomics, 28, 27-39, (1997).

[19] R. Maguire, *"Validating a process for understanding human error probabilities in complex human computer interfaces,"* In, Proceedings of the Second Workshop on Complexity in Design, pp. 81-89, Glasgow University Press, 2005, Glasgow.

[20] E. Lois, V.N. Dang, J. Forester, H. Broberg, S. Massaiu, M. Hildebrandt, P.Ø. Braarud, G. Parry, J. Julius, R. Boring, I. Männistö, and A. Bye, *"International HRA empirical study – Pilot phase report, Description of overall approach and first pilot results from comparing HRA methods to simulator data, HWR-844,"* OECD Halden Reactor Project, 2008, Halden, Norway.

[21] E. Lois, V.N. Dang, J. Forester, H. Broberg, S. Massaiu, M. Hildebrandt, P.Ø. Braarud, G. Parry, J. Julius, R. Boring, I. Männistö, and A. Bye, *"International HRA empirical study—Phase 1 report, Description of overall approach and pilot phase results from comparing HRA methods to simulator performance data, NUREG/IA-0216, Vol. 1,"* US Nuclear Regulatory Commission, 2009, Washington, DC.

[22] A. Bye, E. Lois, V.N. Dang, G. Parry, J. Forester, S. Massaiu, R. Boring, P.Ø. Braarud, H. Broberg, J. Julius, I. Männistö, and P. Nelson, *"The international HRA empirical study—Phase 2 report, Results from comparing HRA method predictions to HAMMLAB simulator data on SGTR scenarios, HWR-915,"* OECD Halden Reactor Project, 2010, Halden, Norway.

[23] J.A. Forester, A.M. Kolaczkowski, V.N. Dang, and E. Lois, *"Human reliability analysis (HRA) in the context of HRA testing with empirical data,"* Official Proceedings of the Joint 8th IEEE Conference on Human Factors and Power Plants and the 13th Annual Workshop on Human Performance/Root Cause/Trending/Operating Experience/Self Assessment, pp. 248-252, IEEE, 2007, New York.

[24] A. Kolaczkowski, J. Forester, E. Lois, and S. Cooper, *"Good practices for implementing human reliability analysis, Final report, NUREG-1792,"* US Nuclear Regulatory Commission, 2005, Washington, DC.

[25] J. Forester, A. Kolaczkowski, S. Cooper, D. Bley, and E. Lois, *"ATHEANA user's guide, NUREG-1880,"* US Nuclear Regulatory Commission, 2007, Washington, DC.

[26] J.A. Forester, E. Lois, V.N. Dang, A. Bye, G. Parry, and J. Julius, *"Lessons learned on human reliability analysis (HRA) methods from the international HRA empirical study,"* Proceedings of the 10th Int. Conf. on Probabilistic Safety Assessment and Management, PSAM10, Paper-362, 2010.

[27] P.Ø. Braarud and B. Johansson, *"Team Cognition in a Complex Accident Scenario, HWR-955,"* OECD Halden Reactor Project, 2010, Halden, Norway.

[28] IEEE, *"Guide for incorporating human action reliability analysis for nuclear power generating stations, IEEE-1082,"* IEEE, 1997, New York.

[29] Electrical Power Research Institute, *"SHARP1—A Revised Systematic Human Action Reliability Procedure, EPRI TR-101711,"* EPRI, 1992, Palo Alto.

[30] H.S. Blackman, D.I. Gertman, and R.L. Boring, *"Human error quantification using performance shaping factors in the SPAR-H method,"* Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society, pp. 1733-1737, (2008).

[31] J. Nielsen and T.K. Landauer, *"A mathematical model of the finding of usability problems,"* Conference on human factors in computing systems: CHI 1993 proceedings, ACM Press, pp. 206-213, (1993).

[32] R.L. Boring, *"Statistical considerations for the number of participants in human factors scaling studies,"* Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society, pp. 1949-1953, (2006).

[33] R.L. Boring and R.L. West, *"Constrained scaling in psychometric magnitude mapping,"* Proceedings of the 24th Annual Meeting of the International Society for Psychophysics, pp. 297-302, (2008).

[34] D. Gertman, H. Blackman, J. Marble, J. Byers, L. Haney, and C. Smith, *"The SPAR-H human reliability analysis method, NUREG/CR-6883,"* US Nuclear Regulatory Commission, 2005, Washington, DC.

[35] V.N. Dang, J.A. Forester, A. Bye, and S. Massaiu, *"Quantitative results of the HRA empirical study and the role of quantitative data in benchmarking,"* Proceedings of the 10th Int. Conf. on Probabilistic Safety Assessment and Management, PSAM10, Paper-350, 2010.

[36] I. Männistö and R. Boring, *"Application of HERA in Empirical HRA Study, HWR-893,"* OECD Halden Reactor Project, 2008, Halden, Norway.

[37] B. Hallbert, R. Boring, D. Gertman, D. Dudenhoeffer, A. Whaley, J. Marble, and J. Joe, *"Human event repository and analysis (HERA) system overview, NUREG/CR-6903, Volume 1,"* US Nuclear Regulatory Commission, 2006, Washington, DC.

[38] R. Boring, A. Whaley, B. Hallbert, K. Laumann, P.Ø Braarud, A. Bye, E. Lois, and Y.H.J. Chang, *"Capturing control room simulator data with the HERA system,"* Official Proceedings of the Joint 8th IEEE Conference on Human Factors and Power Plants and the 13th Annual Workshop on Human Performance/Root Cause/Trending/Operating Experience/Self Assessment, pp. 210-217, IEEE, 2007, New York.